A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search

Jie Tang* Ruoming Jin† Jing Zhang*
* Department of Computer Science and Technology, Tsinghua University, China, 100084
jietang@tsinghua.edu.cn, zhangjing@keg.cs.tsinghua.edu.cn
† Department of Computer Science, Kent State University, Kent, OH 44241
jin@cs.kent.edu

Abstract

In this paper, we propose a unified topic modeling approach and its integration into the random walk framework for academic search. Specifically, we present a topic model for simultaneously modeling papers, authors, and publication venues. We combine the proposed topic model into the random walk framework. Experimental results show that our proposed approach for academic search significantly outperforms the baseline methods of using BM25 and language model, and those of using the existing topic models (including pLSI, LDA, and the AT model).

1 Introduction

Over the last several years, quite a few academic search engines, such as Citeseer, Google Scholar, and Libra, have been built to facilitate the online search over the huge volume of literatures. However, there are still many challenging issues: First, the information-seeking practice [5] is not only about papers, but also about other information sources, such as authors, conferences and journals, etc. Second, academic search typically requires much higher retrieval accuracy. Given a query, such as “data mining”, a user does not typically mean to find papers containing these two words. Her/his intention is to find papers on the data mining topic. Additionally, these two issues are often intertwined.

Existing work has tried to address different aspects of these issues. To rank different objects, random walk over heterogeneous networks has been proposed [9][15]. However, these methods do not consider sub-topics of documents. In the meantime, several work utilizes the topic model to improve the retrieval accuracy [6][14]. However, this is an open issue on how to extend the topic model to deal with heterogeneous data with link information.

In this work, we investigate how the topic model can help with academic search. Specifically, we are interested in:

1. **Heterogeneous topic modeling:** How can we simultaneously model papers, authors, and publication venues within a unified probabilistic topic model?
2. **Academic search:** How to apply the topic model to academic search with better retrieval accuracy?
3. **Ranking with topic model and random walk:** How can we combine the topic model with a random walk framework to improve ranking?

We propose a probabilistic topic model for simultaneously extracting topics of papers, authors, and publication-venues. We further present two methods to combine the proposed topic models with the random walk for ranking different objects simultaneously. Our experimental results show that the proposed method can significantly improve the search quality in comparison with the baseline methods.

2 Preliminary

Let a paper d contain a vector $\mathbf{w}_d$ of N_d words, a vector $\mathbf{a}_d$ of A_d authors, and be published at the venue c_d. Then a collection of D papers can be represented as $\mathbf{D} = \{(\mathbf{w}_1, \mathbf{a}_1, c_1), \cdots, (\mathbf{w}_D, \mathbf{a}_D, c_D)\}$. Table 1 summarizes the notations used in this paper.

We introduce several related work, including: language model [2], pLSI [6], LDA [3], and random walk [10].

Language model is one of the state-of-the-art approaches for information retrieval. It interprets the relevance between a document and a query word as a generative probability:

$$P(w|d) = \frac{N_d}{N_d + \lambda} \cdot \frac{tf(w,d)}{N_d} + (1 - \frac{N_d}{N_d + \lambda}) \cdot \frac{tf(w,\mathbf{D})}{N_{\mathbf{D}}} \quad (1)$$

where $tf(w,d)$ is the word frequency of word w in paper d, $N_{\mathbf{D}}$ is the number of word tokens in the entire collection, and $tf(w,\mathbf{D})$ is the word frequency of word w in the collection $\mathbf{D}$. λ is the Dirichlet smoothing factor. The probability of document d generating a query q can be defined as $P(q|d) = \Pi_{w \in q} P(w|d)$.

Table 1: Notations.

SYMBOL	DESCRIPTION
T	number of topics
D	number of papers
V	number of unique words
A	number of unique authors
C	number of unique publication venues
N_d	number of word tokens in paper d
A_d	number of authors in paper d
$\mathbf{w}_d$	vector form of word tokens in paper d
$\mathbf{a}_d$	vector form of authors in paper d
c_d	the publication venue of paper d
w_{di}	the ith word token in paper d
z_{di}	the topic assigned to word token w_{di}
x_{di}	the chosen author associated with the word token w_{di}
θ_x	multinomial distribution over topics specific to an author x
ϕ_z	multinomial distribution over words specific to topic z
ψ_z	multinomial distribution over publication venues specific to topic z
α, β, μ	Dirichlet priors to multinomial distribution θ, ϕ, and ψ, respectively

Hofmann proposes the probabilistic Latent Semantic Indexing (pLSI) model [6], which assumes that there is a latent topic layer $Z = \{z_1, z_2, \cdots, z_T\}$ between words and documents. Thus, the probability of generating a word w from a document d can be calculated using the topic layer:

$$P(w|d) = \sum_{z=1}^{T} P(w|z)P(z|d) \tag{2}$$

Latent Dirichlet Allocation (LDA) [3] also models documents by using a topic layer. In LDA, for each document d, a multinomial θ_d is first sampled from a Dirichlet with parameter α. Second, for each word w_{di}, a topic z_{di} is chosen from this distribution. Finally, the word w_{di} is generated from a topic-specific multinomial $\phi_{z_{di}}$. Accordingly, the generating probability of word w from document d is:

$$P(w|d, \theta, \phi) = \sum_{z=1}^{T} P(w|z, \phi_z)P(z|d, \theta_d) \tag{3}$$

Several extensions to LDA have been proposed, such as the Author-Topic (AT) model [12].

Considerable research has been conducted on analyzing link structures of the Web, for example PageRank [10] and HITS [7]. Many extensions of the random walk model have been proposed, for example [9] and [15]. Much other effort has also been made for applying the random walk ranking model to mine the bibliography data.

3 Our Proposed Topic Models

Modeling the academic network can be done in many different ways, for example, a separated LDA model for each type of object [14]. However, the separated way may result in unsatisfactory performance. Experimental results in Section 5 confirm this assumption. Our main idea in this work is to use a probabilistic topic model to model papers, authors, and publication venues together.

The proposed model is called Author-Conference-Topic (ACT) model. For simplicity, we use conference to denote all kinds of publication venues, including conference, journal, and article. Essentially, the model utilizes the topic distribution to represent the inter-dependencies among authors, papers, and publication venues. Different strategies can be used to implement the topic model. Specifically, we consider three different types of implementations.

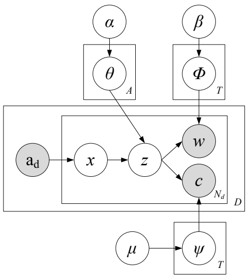

Figure 1: Graphical representation of the ACT1 model.

3.1 ACT Model 1

In the first model, the conference is associated with each word as a stamp. For generating a word w_{di} in paper d, an author x_{di} is first chosen uniformly to be responsible for the word. Each author is associated with a *topic distribution*. Then a topic is sampled from the author-specific topic distribution. Finally the word and the conference stamp is generated from the chosen topic. Figure 1 shows the graphical representation of the ACT model. Formally, the generative process can be described as follows:

1. For each topic z, draw ϕ_z and ψ_z respectively from $Dirichlet(\beta)$ and $Dirichlet(\mu)$;
2. For each word w_{di} in paper d:
    - draw an author x_{di} from $\mathbf{a}_d$ uniformly;
    - draw a topic z_{di} from $Multinomial(\theta_{x_{di}})$ specific to author x_{di}, where θ is generated from $Dirichlet(\alpha)$;
    - draw a word w_{di} from $Multinomial(\phi_{z_{di}})$;
    - draw a conference c_{di} from $Multinomial(\psi_{z_{di}})$.

For inference, the task is to estimate the two sets of unknown parameters in the ACT1 model: (1) the distribution θ of A author-topics, the distribution ϕ of T topic-words, and the distribution ψ of T topic-conferences; and (2) the corresponding topic z_{di} and author x_{di} for each word w_{di}. We use Gibbs sampling for parameter estimation. Instead of estimating the model parameters directly, we calculate the posterior distribution on just x and z and then use the results to infer θ, ϕ, and ψ. The posterior probability is defined as:

$$P(z_{di}, x_{di}|\mathbf{z}_{-di}, \mathbf{x}_{-di}, \mathbf{w}, \mathbf{c}, \alpha, \beta, \mu) \propto$$
$$\frac{m^{-di}_{x_{di}z_{di}} + \alpha_{z_{di}}}{\sum_z (m^{-di}_{x_{di}z} + \alpha_z)} \frac{n^{-di}_{z_{di}w_{di}} + \beta_{w_{di}}}{\sum_v (n^{-di}_{z_{di}v} + \beta_v)} \frac{n^{-d}_{z_{di}c_d} + \mu_{c_d}}{\sum_c (n^{-d}_{z_{di}c} + \mu_c)} \tag{4}$$

where m_{xz} is the number of times that topic z has been associated with the author x, n_{zv} is the number of times that word w_v was generated by topic z, and n_{zc} is the number of times that conference c was generated by topic z; $\mathbf{z}_{-di}$ and $\mathbf{x}_{-di}$ represent all topics and authors assignments excluding the i-th word in the paper d; a number with the superscript

$-di$ denote a quantity, excluding the current instance. α, β, and μ are hyperparameters and were set with fixed values (i.e., $\alpha = 50/T$, $\beta = 0.01$, and $\mu = 0.1$).

During parameter estimation, the algorithm keeps track of a $A \times T$ (author by topic) count matrix, a $T \times V$ (topic by word) count matrix, and a $T \times C$ (topic by conference) count matrix. Given these three count matrices, we can easily estimate the probability of a topic given an author θ_{xz}, the probability of a word given a topic ϕ_{zv}, and the probability of a conference given a topic ψ_{zc}:

$$\theta_{xz} = \frac{m_{xz} + \alpha_z}{\sum_{z'}(m_{xz'} + \alpha_{z'})}, \phi_{zv} = \frac{n_{zv} + \beta_v}{\sum_{v'}(n_{zv'} + \beta_{v'})}$$
$$\psi_{zc} = \frac{n_{zc} + \mu_c}{\sum_{c'}(n_{zc'} + \mu_{c'})} \quad (5)$$

3.2 ACT Model 2 and ACT Model 3

In the second model, each topic is chosen from a multinomial topic distribution specific to an author-conference pair, instead of an author as that in ACT1. The model is derived from the observation: authors usually first choose a publication venue and then write the paper based on themes of the publication venue and interests of the authors. We use a similar method as that in ACT1 for parameter estimation.

In the third model, the conference is taken as a numerical value. Each conference stamp of a paper is sampled after topics have been sampled for all word tokens in the paper. Intuitively, this corresponds to a natural way to publish a paper: authors first write the paper and then determine where to publish it based on topics discussed in the paper. In this model, the conference stamp comes from a normal linear model. Thus, for parameter estimation, there is a slight difference from that in ACT1 and ACT2. We use a Gibbs EM algorithm [1] [8] for inference of this model.

3.3 Applying ACT to Academic Search

Based on the ACT models, we can obtain a form of document model using a similar equation to Equation (3). However, the learned topics by the topic model is usually *general* and not *specific* to a given query. Therefore, only using ACT itself for modeling is too coarse for academic search [14]. Our preliminary experiments also show that employing only the ACT or LDA models to information retrieval hurts the retrieval performance. In general, we would like to have a balance between generality and specificity. Thus, we derive a combination form of the ACT model and the word-based language model:

$$P(w|d) = P_{LM}(w|d) \times P_{ACT}(w|d) \quad (6)$$

where $P_{LM}(w|d)$ is the generating probability of word w from paper d by the language model and $P_{ACT}(w|d)$ is the generating probability by the ACT model.

Similarly, we can define an author model and a conference model in an analogous way:

$$P(q|a) = P_{LM}(q|a) \times P_{ACT}(q|a) \quad (7)$$

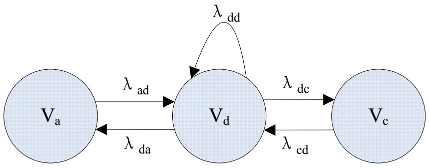

Figure 2: Transition probability

$$P(q|c) = P_{LM}(q|c) \times P_{ACT}(q|c) \quad (8)$$

where in the language model, a and c is represented by a collection of papers published by author a and a collection of papers published on conference c.

4 Ranking with Topic Model and Random Walk

We present two methods to integrate the proposed topic models into the random walk framework.

The academic network is composed of three composite networks (Figure 2). At the center is a directed graph of paper citations $G_d = (V_d, E_{dd})$, where V_d includes all papers, and the directed edge $(d_1, d_2) \in E_{dd}$ suggests the paper d_1 cites the paper d_2. Similarly, relationships between authors and papers are modeled by a bipartite graph $G_{ad} = (V_a \cup V_d, E_{ad})$ and relationships between publication venues and papers are modeled by another bipartite graph $G_{cd} = (V_c \cup V_d, E_{cd})$.

We augment these different graphs to form a heterogeneous graph: $G = (V_d \cup V_a \cup V_c, E_{dd} \cup E_{ad} \cup E_{cd})$. In addition, for the sake of random walk, we represent each (undirected) edge in the bipartite graph as two directed edges, i.e. $\{a_i, d_j\} = (a_i, d_j) \cup (d_j, a_i)$. Further, we define a graph which describes the transition probability between different types of nodes. Clearly, we need $\lambda_{dd} + \lambda_{da} + \lambda_{dc} = 1$. We also define $\lambda_{ad} = \lambda_{cd} = 1$. This transition graph formalizes a random surfer's behavior as follows. The random surfer will have the λ_{dd} probability to stay in the paper citation network, and will have λ_{da} and λ_{dc} probabilities to find authors and the publication venue related to the paper.

Given this, similar to PageRank, we can define a general form of the random walk ranking score for each node x as:

$$r[x] = \frac{\xi}{|V|} + (1 - \xi) \times \sum_{(x,y) \in E} \lambda_{yx} r[y] P(x|y) \quad (9)$$

where $|V|$ is the number of nodes in the network; ξ is a random jump parameter; λ_{yx} is the transition probability between the type of node y and the type of node x; $P(x|y)$ is the probability between two specific nodes y and x.

4.1 Combination Method 1

The first method combines the random walk ranking score with the relevance score from the topic model by simple multiplication. Formally, given a query q, the final ranking score of a paper d is the multiplication of random walk

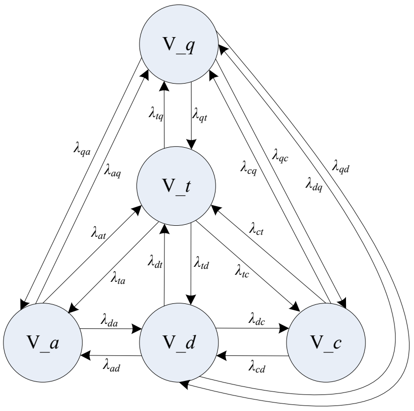

Figure 3: Transition probability

ranking score $r[d]$ and the paper's relevance score $P(q|d)$:

$$R[d] = r[d] \times P(q|d) \qquad (10)$$

Similarly, we can define the combination formulas for conferences and authors:

$$R[a] = r[a] \times P(q|a), \quad R[c] = r[c] \times P(q|c) \qquad (11)$$

We also tried the weighted sum method for combination, i.e. $R[d] = \gamma \times r[d] + (1 - \gamma) \times P(q|d)$, where γ is a coefficient to control the strength of the two terms. It always underperforms the method of the multiplication combination, even with different values of the coefficient γ.

4.2 Combination Method 2

The second method directly integrates the topic model into the random walk. It augments the network with topic nodes V_t and a query node V_q.

Let $G_{td} = (V_t \cup V_d, E_{td})$ be a bipartite graph where V_t is the set of topic nodes estimated in the topic model, and if paper d can be generated from topic z with a probability $P(\mathbf{w}_d|z) > \epsilon$ (where ϵ is a parameter to control the density of the constructed network), then we have an edge $(z, d) \in E_{td}$. Similarly, we can define edges E_{ct} between conferences and topics and edges E_{at} between authors and topics. Furthermore, we add edges between the query node and different other nodes (papers, authors, conferences, and topics). Figure 3 shows the new heterogeneous network.

In this method, we consider that after the random surfer walks to a topic node from some other node, he/she will always walk to the (virtual) query node and then walk back to another topic node. The transition probability between the query and a paper/author/conference node is calculated using the language model (e.g., $P_{LM}(q|d)$). The transition probabilities between a topic node and the other nodes are calculated using the topic model, i.e.,

$$P(z_i|a_j) = \theta_{a_j z_i}, \quad P(a_j|z_i) = \frac{P(z_i|a_j)P(a_j)}{P(z_i)}$$

$$P(z_i|d_j) = \frac{1}{A_d} \sum_{x \in \mathbf{a}_d} \theta_{x z_i}, \quad P(d_j|z_i) = \prod_{i=1}^{N_d} P(w_{di}|z_i)$$

$$P(c_j|z_i) = \psi_{z_i c_j}, \quad P(z_i|c_j) = \frac{P(c_j|z_i)P(z_i)}{P(c_j)}$$

$$P(q|z_i) = \prod_{w \in q} P(w|z_i), \quad P(z_i|q) \propto P(q|z_i)P(z_i) \qquad (12)$$

where θ and ψ are obtained from (5); $P(z_i)$, $P(a_j)$, and $P(c_j)$ can be obtained by counting the number after Gibbs sampling.

The method can make use of the topic distribution in the random walk and we can also adjust the different λ between the other nodes to the topic nodes to weight how the random walk and the topic model affect the final rank. The ranking scores after random walk are query dependent.

5 Experimental Results

We evaluated the proposed models in our developed system ArnetMiner (http://arnetminer.org) [13].

5.1 Experimental Setting

5.1.1 Data sets

As there is no a standard data set with ground truth and also it is difficult to create such a data set of ground truth, for evaluation purpose, we collected 43 most frequent queries from the query log of ArnetMiner. We divided these queries into two sub sets and carried out two experiments.

In the first experiment, we used 7 queries and conducted evaluation on a subset of the data (including $14,134$ authors, $10,716$ papers, and $1,434$ conference) from ArnetMiner. For evaluation, we used the method of pooled relevance judgments [4] together with human judgments. Specifically, for each query, we first pooled the top 30 results from three academic search systems (Libra, Rexa, and ArnetMiner). Then, two faculties and five graduates from CS provided human judgments. Four grade scores (3, 2, 1, and 0) were assigned respectively representing best relevance, relevance, marginal relevance, and not relevance.

In the second experiment, we used the rest 36 queries and conducted evaluation on the entire data of ArnetMiner (including $448,365$ authors, $981,599$ papers, and $4,501$ conferences). For evaluation, we used only the pooled relevance judgments without human judgements. For pooling purpose, we added one baseline method, i.e. BM25 [11]. Specifically, for a query, we pooled the top 30 results returned by BM25, our method, and two systems: Libra and Rexa. We define as relevance only when a candidate was returned by at least three of the four methods.

5.1.2 Experimental Setting

In all experiments, we conducted evaluation in terms of P@5, P@10, P@20, R-pre, and MAP [4].

We used BM25 [11], language model (LM) [2], pLSI [6], LDA [3], and the Author-Topic (AT) model [12] as baseline methods. BM25 is a state-of-the-art method for information retrieval. In BM25, we used the method in [11] to calculate the relevance of a query and a paper. For language model, we used Equation (1) to calculate the relevance between a query term and a paper and for pLSI, we used Equation (2) to calculate the relevance. For LDA, we used Equation (3) to calculate the relevance of a term and a paper. For the AT model, we used similar equations to Equation (6) and (7) to calculate the relevance of a query term with a paper and an author respectively. We also compared with the results obtained by combining LM with random walk using combination method 1. We also tried to combine BM25 with RW1. The result underperforms that of LM+RW1.

To learn the pLSI model, we used the EM algorithm [6]. For LDA and AT, we performed model estimation with the same setting as that for the ACT models. In the first experiment, we empirically set the number of topics as $T = 80$ for all topic models. In the second experiment, we set the number of topics as $T = 200$.

5.2 Experimental Results

5.2.1 Performances of the First Experiment

Table 2 shows the performance of retrieving papers, conferences, and authors using our proposed methods and the baseline methods. +RW denotes integration of a method into the random walk. RW1 denotes the combination method 1, and RW2 denotes the combination method 2. We see that our proposed topic models outperform all the baseline methods (BM25, LM, pLSI, LDA, and AT). We can also see that ACT1+RW1 achieves the best performance in terms of all evaluation measures.

Table 2: Performance of academic ranking approaches (%).

Method	Object	P@5	P@10	P@20	R-pre	MAP
BM25	Paper	42.9	45.7	41.4	12.0	47.2
	Author	77.1	47.1	26.4	67.5	85.5
	Conference	51.4	38.6	22.9	48.8	66.0
	Average	57.1	43.8	30.2	42.8	66.2
LM	Paper	40.0	38.6	37.1	10.0	46.4
	Author	65.7	44.3	25.0	58.8	73.4
	Conference	51.4	32.9	21.4	47.6	63.1
	Average	52.4	38.6	27.9	38.8	61.0
LM+RW1	Paper	62.9	55.7	44.3	12.9	65.3
	Author	71.4	48.6	25.7	64.6	83.8
	Conference	60.0	35.7	22.1	53.6	64.6
	Average	64.8	46.7	30.7	43.7	71.2
pLSI	Paper	32.5	33.8	30	9.7	40.4
	Author	65.0	40.0	22.5	60.4	75.5
	Conference	47.5	36.3	21.3	45.1	54.1
	Average	48.3	36.7	24.6	38.4	56.7
LDA	Paper	31.4	48.6	42.9	13.5	45.8
AT	Paper	42.9	48.6	42.9	13.1	49.3
	Author	82.9	45.7	25.7	73.5	78.1
	Average	62.9	47.1	34.3	43.3	63.7
ACT1	Paper	42.9	45.7	43.6	16.6	51.0
	Author	91.4	50.0	26.4	80.0	89.6
	Conference	62.9	41.4	23.6	60.7	72.3
	Average	65.7	45.7	31.2	52.4	71.0
ACT1+RW1	Paper	68.6	61.4	50.7	17.1	66.6
	Author	80.0	51.4	27.1	77.6	87.4
	Conference	62.9	42.9	23.6	59.5	72.0
	Average	**70.5**	**51.9**	**33.8**	**51.4**	**75.4**
ACT1+RW2	Paper	45.7	40.0	38.6	13.4	52.2
	Author	71.4	44.3	24.3	65.4	71.5
	Conference	51.4	32.9	20.0	53.6	60.7
	Average	56.2	39.1	27.6	44.1	61.4
ACT2	Paper	42.9	47.1	39.3	15.0	47.7
	Author	74.3	50.0	25.7	69.4	80.1
	Conference	54.3	41.4	22.1	54.2	63.9
	Average	57.1	46.2	29.1	46.2	63.9
ACT2+RW1	Paper	62.9	58.6	48.6	16.9	63.5
	Author	77.1	50.0	26.4	69.4	82.8
	Conference	60.0	40.0	23.6	55.4	65.8
	Average	66.7	49.5	32.9	47.2	70.7
ACT2+RW2	Paper	40.0	41.4	40.0	13.5	47.3
	Author	74.3	42.9	24.3	63.6	72.2
	Conference	51.4	32.9	22.1	50.0	63.1
	Average	55.2	39.1	28.8	42.4	60.9
ACT3	Paper	42.9	38.6	41.4	17.1	47.0
	Author	71.4	47.1	25.7	70.0	78.7
	Conference	57.1	38.6	23.6	58.3	65.7
	Average	57.1	41.4	30.2	48.5	63.8
ACT3+RW1	Paper	65.7	54.3	49.3	19.5	63.0
	Author	77.1	50.0	26.4	74.8	84.0
	Conference	60.0	41.4	23.6	55.4	66.0
	Average	67.6	48.6	33.1	49.9	71.0
ACT3+RW2	Paper	40.0	38.6	38.6	14.0	47.4
	Author	74.3	41.4	24.3	63.6	72.0
	Conference	54.3	32.9	22.1	50.0	62.6
	Average	56.2	37.6	28.3	42.5	60.7

5.2.2 Performances of the Second Experiment

In the second experiment, we evaluated the results of our best approach (ACT1+RW1), one baseline method: BM25, and the two systems: Libra and Rexa. (Rexa only has paper and author search.) Table 3 shows the performance of the four methods. We see that our proposed method outperforms the baseline method and the two systems.

Table 3: Performance of academic ranking approaches (%).

Method	Object	P@5	P@10	P@20	R-pre	MAP
ACT1+RW1	Paper	39.3	35.7	31.6	30.9	46.7
	Author	45.7	37.9	30.7	33.6	50.6
	Conference	48.6	41.8	30.9	47.4	52.9
	Average	44.5	38.5	31.1	37.3	50.1
BM25	Paper	41.4	35.0	33.2	31.7	49.1
	Author	43.6	33.2	29.3	27.9	51.9
	Conference	41.4	33.9	28.9	36.9	46.0
	Average	42.1	34.0	30.5	32.2	49.0
Libra	Paper	35.0	26.1	24.1	23.2	38.0
	Author	37.1	30.7	27.9	21.8	44.8
	Conference	48.6	37.5	29.5	47.3	57.4
	Average	40.2	31.4	27.1	30.8	46.7
Rexa	Paper	27.9	21.1	15.4	20.9	38.9
	Author	24.3	19.6	15.9	17.3	30.1
	Average	26.1	20.4	15.6	19.1	34.5

5.3 Discussion

(1) Our proposed methods for academic search significantly outperform the baseline methods. We see from both Table 2 and Table 3 that ACT1+RW1 achieves the best performance. This indicates that the proposed approach indeed improves the quality of academic search.

(2) We see that by combining the topic models with random walk, we can significantly enhance the ranking perfor-

mance (+4.4%, +6.8%, and +7.2% respectively for ACT1, ACT2, and ACT3 in terms of MAP).

We have also observed a surprising result: the first combination method uses the simple multiplication to combine the relevance scores by the topic model with the score from the random walking model while the second method integrates the topic model directly into the random walk. Intuitively, the second combination method seems to work in a more "elegant" way and would have resulted in a better performance. However, results in Table 2 show that the first simple combination method obtains significant improvements while the second combination method hurts the retrieval performance (e.g., -9.6% for ACT1 when combining random walk using the method 2).

Further analysis shows that the second combination method integrates the topic model into each step in the random walk, which leads to too many parameters to tune. For example, there are eighteen λ needed to tune (cf. Figure 3). It is hard to find a best setting for all the parameters.

(3) We can also see from Table 2 that the proposed different topic models perform different behaviors, but ACT1 performs the best either with (+4.7% than ACT2 and +4.4% than ACT3 in terms of MAP) or without the random walk (+7.1% than ACT2 and +7.2% than ACT3 by MAP).

With a further analysis, we found that the sampling process of ACT2 results in a very huge and sparse (author-conference-pair by topic) count matrix of $AC \times T$, which is scaled up to 10million rows. The problem of ACT3 might be that we take the conference stamp as a numerical value, which makes it not accurate enough to describe the conference information of discrete value. How to accurately capture the conference information in the topic model is also one of our ongoing research issues.

(4) Experiments (cf. Table 3) show that our approach outperforms the existing academic search systems Libra (+3.4% in terms of average MAP) and Rexa (+15.6% by average MAP). The reasons include: (1) Rexa only supports paper search and author search, thus cannot make use of dependencies between authors, papers, and conferences; (2) Libra supports search of the three objects. However, its ranking is based on the language model and the PageRank method [9]. Again, it cannot make use of the semantic (topic-level) dependencies between different objects.

6 CONCLUSION

In this paper, we investigate the problem of modeling heterogeneous academic network using a unified probabilistic model. We present three topic models for simultaneously modeling papers, authors, and publication venues. We further propose two methods to combine the proposed topic models with the random walk framework for academic search. Experimental results show that our approach outperforms the baseline methods and the existing systems.

Our proposed approach is very general and flexible. The topic model can be implemented in many other ways and the random walk can run in either an offline mode or an online mode. Variations of the approach can be applied to many other applications such as social search and blog search.

7 ACKNOWLEDGMENTS

Jie Tang and Jing Zhang would like to thank the support of NSFC research funding (90604025, 60703059), Chinese 973 research funding (2007CB310803), and Chinese Young Faculty Funding (20070003093).

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50:5–43, 2003.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of SIGIR'04*, pages 25–32.

[5] M. Hertzum and A. M. Pejtersen. The information-seeking practices of engineers: Searching for documents as well as for people. *Information Processing & Management*, 36(5):761–778, 2000.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of SIGIR'99*, pages 50–57, 1999.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[8] T. Minka. Estimating a dirichlet distribution. Technical report, http://research.microsoft.com/ minka/papers/dirichlet/, 2003.

[9] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *Proc. of WWW'05*, pages 567–574, 2005.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

[11] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In *TREC'96*, 1996.

[12] M. Steyvers, P. Smyth, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. of KDD'04*.

[13] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proc. of SIGKDD'08*, pages 990–998, 2008.

[14] X. Wei and W. B. Croft. Lda-based document models for adhoc retrieval. In *Proc. of SIGIR'06*, pages 178–185, 2006.

[15] W. Xi, B. Zhang, Y. Lu, Z. Chen, S. Yan, H. Zeng, W.-Y. Ma, and E. A. Fox. Link fusion: a unified link analysis framework for multi-type interrelated data objects. In *Proc. of WWW'04*, pages 319–327, 2004.